

HUBUNGAN FUNGSI SIMILARITAS DAN PERSEPSI PENUTUR DALAM MENENTUKAN SKOR SIMILARITAS TEKS BAHASA INDONESIA



**Disusun sebagai salah satu syarat menyelesaikan Program Studi Strata I pada Jurusan
Informatika Fakultas Komunikasi dan Informatika**

Oleh:

ARIF NUGROHO PUTRO

L200140147

**PROGRAM STUDI INFORMATIKA
FAKULTAS KOMUNIKASI DAN INFORMATIKA
UNIVERSITAS MUHAMMADIYAH SURAKARTA
2018**

HALAMAN PERSETUJUAN

**HUBUNGAN FUNGSI SIMILARITAS DAN PERSEPSI PENUTUR
DALAM MENENTUKAN SKOR SIMILARITAS TEKS BAHASA
INDONESIA**

PUBLIKASI ILMIAH

oleh:

ARIF NUGROHO PUTRO

L200140147

Telah diperiksa dan disetujui untuk diuji oleh:

Dosen Pembimbing



Husni Thamrin, S.T., M.T., Ph.D.

NIK. 970

HALAMAN PENGESAHAN

HUBUNGAN FUNGSI SIMILARITAS DAN PERSEPSI PENUTUR DALAM MENENTUKAN SKOR SIMILARITAS TEKS BAHASA INDONESIA

OLEH

ARIF NUGROHO PUTRO

L200140147

Telah dipertahankan di depan Dewan Penguji
Fakultas Komunikasi dan Informatika
Universitas Muhammadiyah Surakarta
Pada hari 03, Agustus 2017
dan dinyatakan telah memenuhi syarat

Dewan Penguji:

1. Husni Thamrin, S.T., M.T., Ph.D.
(Ketua Dewan Penguji)
2. Aris Rakhmadi, S.T., M.T.
(Anggota I Dewan Penguji)
3. Dr. Heru Supriyono, M.Sc.
(Anggota II Dewan Penguji)

(.....)
(.....)
(.....)

Dekan

Fakultas Komunikasi dan Informatika



Nurgiyatna, S.T., M.Sc., Ph.D.

NIK. 123

Ketua

Program Studi Informatika



Heru Supriyono, S.T., M.Sc.

NIK. 970

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam naskah publikasi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali secara tertulis diacu dalam naskah dan disebutkan dalam daftar pustaka.

Apabila kelak terbukti ada ketidakbenaran dalam pernyataan saya di atas, maka akan saya pertanggungjawabkan sepenuhnya.

Surakarta, 03 Agustus 2018

Penulis



Arif Nugroho Putro

L200140147



**UNIVERSITAS MUHAMMADIYAH SURAKARTA
FAKULTAS KOMUNIKASI DAN INFORMATIKA
PROGRAM STUDI INFORMATIKA**

Jl. A Yani Tromol Pos 1 Pabelan Kartasura Telp. (0271)717417, 719483 Fax (0271) 714448
Surakarta 57102 Indonesia Web: <http://informatika.ums.ac.id> Email: informatika@ums.ac.id

SURAT KETERANGAN LULUS PLAGIASI

NO SURAT 210/A.3-R.3/INT-FKI/VM/2018

Assalamu'alaikum Wr. Wb

Biro Skripsi Program Studi Informatika menerangkan bahwa :

Nama : ARIF NUGROHO PUTRO
NIM : L200140147
Judul : Hubungan Fungsi Similaritas dan Persepsi Penutur dalam
menentukan Skor Similaritas teks Bahasa Indonesia
Program Studi : INFORMATIKA
Status : Lulus


Adalah benar-benar sudah lulus pengecekan plagiasi dari Naskah Publikasi Skripsi, dengan menggunakan aplikasi Turnitin.

Demikian surat keterangan ini dibuat agar dipergunakan sebagaimana mestinya.

Wassalamu'alaikum Wr. Wb

Surakarta, 31 Juli 2018

Biro Skripsi Informatika


Ihsan Cahyo Utomo, S.Kom., M.Kom.



UNIVERSITAS MUHAMMADIYAH SURAKARTA
FAKULTAS KOMUNIKASI DAN INFORMATIKA
PROGRAM STUDI INFORMATIKA

Jl. A Yani Tromol Pos 1 Pabelan Kartasura Telp. (0271)717417, 719483 Fax (0271) 714448
Surakarta 57102 Indonesia. Web: <http://informatika.ums.ac.id> Email: informatika@ums.ac.id

Secure | https://eui.tuttnin.com/acc/carta/en_us/?c=9855141138Rs=1&lang=en_us&u=105/550000

feedback studio

HUBUNGAN FUNGSI SIMILARITAS DAN PERSEPSI PENUTUR DALAM MENENTUKAN SKOR SIMILARITAS TEKS BAHASA INDONESIA

HUBUNGAN FUNGSI SIMILARITAS DAN PERSEPSI PENUTUR
DALAM MENENTUKAN SKOR SIMILARITAS TEKS BAHASA INDONESIA

Arif Nugroho Patro, Husni Thamrin

Abstrak

Algoritma similaritas banyak diterapkan diberbagai bidang seperti klasifikasi dokumen, pencarian dokumen, deteksi plagiasi, sistem koreksi otomatis dan sistem tanya jawab otomatis. Penelitian ini bertujuan untuk membandingkan hasil penilaian manusia dengan hasil perhitungan skor fungsi similaritas terhadap klenipatah tutan sepuang lalinat. Untuk mendapatkan skor persepsi manusia peneliti membuat website survei untuk membandingkan sebuah pasangan kalimat. Peneliti juga menggunakan tiga fungsi similaritas yaitu cosine similarity, dice similarity dan euclidean distance. Menurut hasil perhitungan antara persepsi manusia dan tiga fungsi similaritas, skor kesamaan fungsi cosine similarity adalah 0.680303, dice similarity 0.684848, sedangkan euclidean distance 0.348485 dari ketiga hasil perhitungan similaritas tersebut, peneliti menyimpulkan bahwa dice similarity merupakan fungsi similaritas yang terbaik untuk menghitung tingkat similaritas sebuah pasangan kalimat, sementara euclidean distance kurang baik untuk menghitung similaritas pasangan kalimat.

Kata Kunci: similaritas, cosine, cosine similarity, dice similarity, euclidean distance.

Abstract

Algorithm similarity widely applied in various fields such as document classification, documents search, plagiarism detection, automatic correction system and automatic question and answer system. This study aims to compare the results of human appraisal with the result of calculation / score of similarity function to the similarity between a pair of sentences. To get a human perception score the researcher makes a survey website to compare a pair of sentences. The researcher also used three similarity functions: cosine similarity

Match Overview

10%

- | Rank | Source | Score |
|------|-------------------------|-------|
| 1 | publikasi.ums.ac.id | 2% |
| 2 | egwits.ums.ac.id | 2% |
| 3 | Submitted to University | 2% |
| 4 | Submitted to University | 1% |
| 5 | journal.ums.ac.id | 1% |
| 6 | doi.org | 1% |
| 7 | Mohamed Binta, Rachid | <1% |

Page 2 of 14

Word Count: 2891

Text-only Report

High Resolution

On

HUBUNGAN FUNGSI SIMILARITAS DAN PERSEPSI PENUTUR DALAM MENENTUKAN SKOR SIMILARITAS TEKS BAHASA INDONESIA

Abstrak

Algoritma similaritas banyak diterapkan diberbagai bidang seperti klasifikasi dokumen, pencarian dokumen, deteksi plagiasi, sistem koreksi otomatis dan sistem tanya jawab otomatis. Penelitian ini bertujuan untuk membandingkan hasil penilaian manusia dengan hasil perhitungan/skor fungsi similaritas terhadap kemiripan antara sepasang kalimat. Untuk mendapatkan skor persepsi manusia peneliti membuat website survei untuk membandingkan sebuah pasangan kalimat. Peneliti juga menggunakan tiga fungsi similaritas yaitu *cosine similarity*, *dice similarity* dan *euclidean distance*. Menurut hasil perhitungan antara persepsi manusia dan tiga fungsi similaritas, skor kesamaan fungsi *cosine similarity* adalah 0.680303, *dice similarity* 0.684848. sedangkan *euclidean distance* 0.348485. dari ketiga hasil perhitungan similaritas tersebut, peneliti menyimpulkan bahwa *dice similarity* merupakan fungsi similaritas yang terbaik untuk menghitung tingkat similaritas sebuah pasangan kalimat, sementara *euclidean distance* kurang baik untuk menghitung similaritas pasangan kalimat.

Kata Kunci: similaritas, survei, *cosine similarity*, *dice similarity*, *euclidean distance*.

Abstract

Algorithm similarity widely applied in various fields such as document classification, document search, plagiarism detection, automatic correction system and automated question and answer system. This study aims to compare the results of human appraisal with the result of calculation / score of similarity function to the similarity between a pair of sentences. To get a human perception score the researcher makes a survey website to compare a pair of sentences. The researchers also used three similarity functions: cosine similarity, dice similarity and euclidean distance. According to the calculation result between human perception and three similarity function, cosine similarity function similarity score is 0.680303, dice similarity 0.684848. while euclidean distance 0.348485. of the three similarity calculations, the researchers conclude that the dice similarity is the best similarity function to calculate the similarity level of a sentence pair, while the euclidean distance is less good to calculate the similarity of the sentence pairs.

Keywords: similarity, survey, *cosine similarity*, *dice similarity*, *euclidean distance*.

1. PENDAHULUAN

Algoritma similaritas banyak diterapkan diberbagai bidang seperti klasifikasi dokumen (Wahyuni, 2017), pencarian dokumen (Putung, 2016), deteksi plagiasi (Leonardo & Hansun, 2017), sistem koreksi otomatis (Thamrin & Wantoro, 2014) dan sistem tanya jawab otomatis (Syauqi & Nurwahdah, 2015). Pada proses klasifikasi dokumen, dokumen teks dicermati kontennya untuk diberi label atau kategori berdasarkan similaritas konten dokumen standar dengan konten dokumen yang hendak diklasifikasikan. Pada proses pencarian dokumen, kesamaan beberapa dokumen akan dibandingkan

dengan parameter kata kunci sebagai perbandingannya, setelah itu dokumen dianalisa untuk melakukan pengecekan antar dokumen. Pada proses deteksi plagiasi, isi dari suatu dokumen yang diuji akan dibandingkan dengan beberapa dokumen lainnya dengan menghitung tingkat persamaan dokumen agar dapat terdeteksi apakah dokumen yang dibuat berupa plagiat atau tidak. Pada sistem tanya jawab otomatis, kalimat yang ditanyakan akan dicocokkan kemiripannya dari pertanyaan yang sudah tersedia di dalam penampungan kata sehingga akan memberikan jawaban yang sesuai.

Ada berbagai macam algoritma similaritas yang dapat digunakan seperti algoritma TF-IDF, algoritma Winnowing, algoritma Rabin-Karp, dan Algoritma Semantik (Thamrin & Sabardilla, 2015). Algoritma TF-IDF (*Term Frequency Inverse Document Frequency*) merupakan suatu metode kombinasi dari dua konsep. TF (*Term Frequency*) sebagai frekuensi jumlah nilai kemunculan suatu kata yang ada pada dokumen, jika kemunculan kata yang diuji semakin banyak maka nilainya akan semakin tinggi. IDF(*Inverse Document Frequency*) yang menyatakan banyaknya kata pada kalimat yang muncul pada dokumen yang nantinya akan dihitung dengan mengkalikan TF dan IDF. Algoritma Winnowing merupakan suatu metode yang dapat mengidentifikasikan kesamaan suatu kata melalui algoritma rolling hash, yang menggunakan perhitungan ASCII setiap karakter dan menghilangkan tanda baca. Prinsip Algoritma Rabin-Karp yaitu dengan menggunakan fungsi hash dan mencari kata berupa substring dalam suatu dokumen. Algoritma Semantik tertuju pada arti kata dari kalimat yang terbentuk.

Algoritma similaritas dapat digolongkan menjadi dua yaitu algoritma yang bersifat *language dependent* (tergantung bahasa) dan *language independent* (tidak tergantung bahasa). Algoritma yang tergantung bahasa membutuhkan pengetahuan tentang struktur bahasa, tata bahasa dan kumpulan kata atau kalimat dalam bahasa yang dipakai. Contoh algoritma yang tergantung bahasa adalah algoritma yang memanfaatkan kamus dan wordnet (Zamzami, 2016). Algoritma yang tidak tergantung bahasa hanya memerlukan kumpulan kosa kata atau kalimat dan tidak memerlukan pengetahuan tentang makna kata atau tata bahasa. Contohnya adalah algoritma TF-IDF (Wahyuni, 2017).

Namun apapun jenis algoritma yang digunakan, selalu diperlukan pengujian apakah algoritma yang diterapkan sesuai dengan apa yang dipahami oleh manusia. Penelitian ini membutuhkan data survei untuk menyelesaikan permasalahan menghitung similaritas kalimat, kemudian data survei diolah untuk pengujian. Untuk mendapatkan data survei dari seseorang diperlukan aplikasi survei berbasis website. Selain menggunakan data survei dari seseorang, beberapa fungsi similaritas juga digunakan untuk menentukan nilai similaritas, yaitu *cosine similarity*, *dice similarity* dan *euclidean distance*. Penelitian ini berupaya mencermati sejauh mana hubungan beberapa algoritma fungsi

similaritas dengan persepsi orang indonesia terkait similaritas sebuah pasangan kalimat. Sehingga mendapatkan hasil perhitungan similaritas dari persepsi seseorang dengan fungsi similaritas.

2. METODE

Penelitian ini dilakukan untuk menilai persepsi penutur seseorang terhadap sebuah pasangan kalimat, kemudian penilaian tersebut akan dibandingkan dengan perhitungan fungsi similaritas. Fungsi similaritas yang dipilih pada penelitian ini adalah fungsi *cosine similarity*, *dice similarity* dan *euclidean distance* (Hamzah, A., 2008). Berikut rumus dari fungsi similaritas :

1. Cosine similarity

$$Cos = \frac{\sum (D1 \times D2)}{\sqrt{\sum (D1)^2 \times \sum (D2)^2}} \quad (1)$$

2. Dice similarity

$$Dice = \frac{2 \times \sum (D1 \times D2)}{\sum (D1) + \sum (D2)} \quad (2)$$

3. Euclidean distance

$$Euclidean = \sqrt{\sum (D1 - D2)^2} \quad (3)$$

Keterangan :

\sum = Jumlah semua data

D1 = Variabel Kalimat Pertama, yang akan dibandingkan kemiripannya.

D2 = Variabel Kalimat Kedua, yang akan dibandingkan kemiripannya.

Untuk menghitung pasangan kalimat menggunakan fungsi similaritas, yang dilakukan pertama adalah menggabungkan pasangan kalimat yang nantinya akan dijadikan sebagai perbandingan terhadap pasangan kalimat, lalu memilih kata yang berbeda, jika ada kata yang sama ditulis satu saja dan disimpan pada suatu variabel D3. Setelah disimpan, pasangan kalimat D1 dan D2 dibandingkan dengan variabel D3, jika pasangan kalimat terdapat kata yang sama di variabel D3 maka akan bernilai 1, jika tidak ada akan bernilai 0. Dari perbandingan kalimat D1 dan D2 akan didapat angka yang dapat dihitung menggunakan rumus fungsi similaritas.

Survei digunakan pada penelitian ini untuk pengumpulan data sebuah pasangan kalimat yang dilakukan menggunakan website, data yang dikumpulkan berupa pendapat dari setiap orang tentang similaritas kalimat yang dibandingkan dengan memberikan nilai pada kalimat. Kemudian pasangan kalimat dihitung dengan fungsi similaritas berdasarkan persamaan kata pada penyusunan kalimat, pada proses penghitungan antara persepsi seseorang dan fungsi similaritas akan didapatkan skor yang menjadi tujuan dari penelitian ini.

Data yang diuji untuk penelitian ini berupa 1000 pasangan kalimat yang ditujukan pada setiap orang khususnya mahasiswa, pemberian nilai dilakukan dengan memilih skor dari pilihan yang sudah disediakan website. Pilihan yang disediakan untuk memberikan skor pada pasangan kalimat yaitu mirip dan tidak mirip. langkah penelitian ini dimulai dengan mengumpulkan data survei dari penilaian seseorang, penilaian dilakukan hingga 1000 pasangan kalimat selesai, lalu diulangi dengan responden yang berbeda. kemudian data survei diseleksi berdasarkan penilaian seseorang yang sama.

Dalam menghitung skor menggunakan fungsi similaritas, setiap pasangan kalimat yang dibandingkan hanya menggunakan gabungan kata dari kalimat 1 dan kalimat 2 untuk mendapatkan nilai persamaan dari pasangan kalimat. Pada penelitian ini, data yang diuji tidak melibatkan perbandingan kata negasi seperti “tidak”, “bukan”, “tanpa”, “tiada” dan sebagainya. Setelah mendapatkan skor dari setiap fungsi, skor tersebut dibandingkan berdasarkan ketentuan setiap fungsi untuk mendapatkan penilaian antara mirip atau tidak mirip. Ketentuan perbandingan untuk *cosine similarity* dan *dice similarity* yaitu jika skor yang dihitung menggunakan rumus dasar lebih dari atau sama dengan 0.5 maka penilaian dikatakan mirip dan jika skor kurang dari 0.5 maka penilaian dikatakan tidak mirip, sedangkan untuk *euclidean distance* jika skor yang dihitung menggunakan rumus dasar kurang dari atau sama dengan 2 maka penilaian dikatakan mirip dan jika skor lebih dari 2 maka penilaian dikatakan tidak mirip.

Kemudian hasil seleksi data survei berdasarkan penilaian yang sama akan dibandingkan dengan hasil perbandingan berdasarkan ketentuan setiap fungsi, perbandingan antara keduanya memiliki ketentuan yaitu jika hasil data survei sama dengan hasil fungsi maka dinilai 1 yang berarti mempunyai kesamaan antara persepsi manusia dengan fungsi similaritas dan jika hasil data survei tidak sama dengan hasil fungsi maka dinilai 0 yang berarti tidak memiliki kesamaan. Setelah mendapatkan hasil kesamaan antara persepsi manusia dengan fungsi similaritas, semua hasil dari setiap fungsi dijumlahkan dan hasil dari penjumlahan tersebut akan dibagi dengan seluruh data survei yang sudah diseleksi tadi. Sehingga mendapatkan tingkat kesamaan dari setiap fungsi similaritas.

Alat dan bahan yang digunakan dalam penelitian ini berupa laptop dan perangkat lunak, perangkat yang digunakan dalam membuat web ini adalah sistem operasi Microsoft Windows 7 Ultimate, Notepad++, Bahasa pemrograman PHP, webserver Apache dan Server basis data mysql 5.

Cara kerja website :

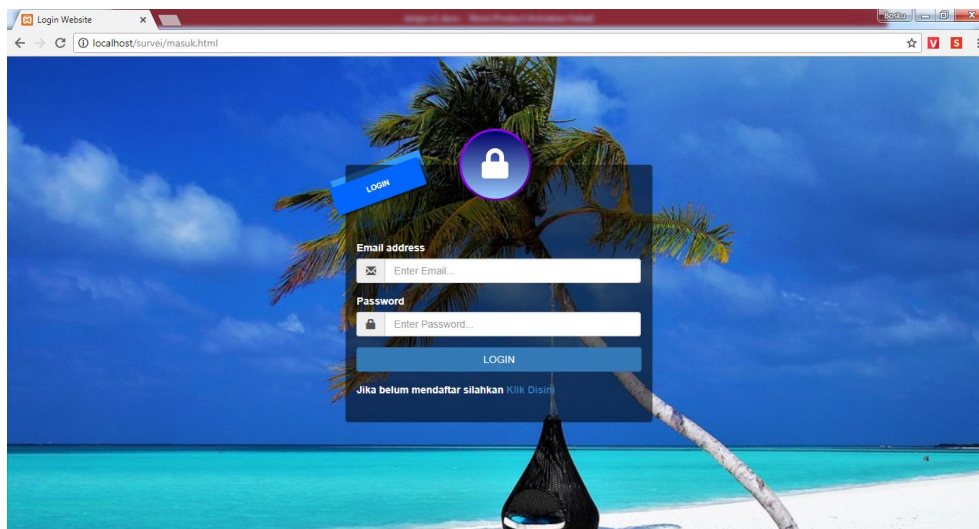
- a. User membuka website untuk login.
- b. User diberikan survei terkait pasangan-pasangan kalimat yang akan dibandingkan.
- c. User memberikan nilai pada setiap pasangan kalimat yang dibandingkan.
- d. User menekan tombol selesai untuk mengakhiri survei.
- e. Website menampilkan ucapan terima kasih.

3. HASIL DAN PEMBAHASAN

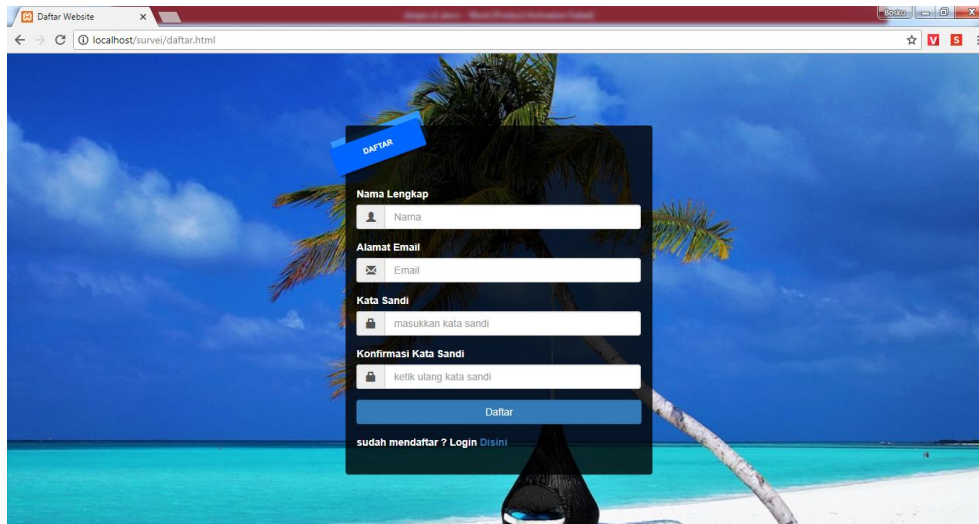
3.1 TAMPILAN SISTEM

3.1.1 HALAMAN LOGIN & PENDAFTARAN

Gambar 1 merupakan halaman login ketika responden mengakses website ini. terdapat form email dan form password untuk login ke menu utama. jika responden belum pernah mendaftar, pada gambar 2 merupakan halaman pendaftaran agar responden dapat login ke website ini. terdapat form nama lengkap , form email , form password dan form konfirmasi password.



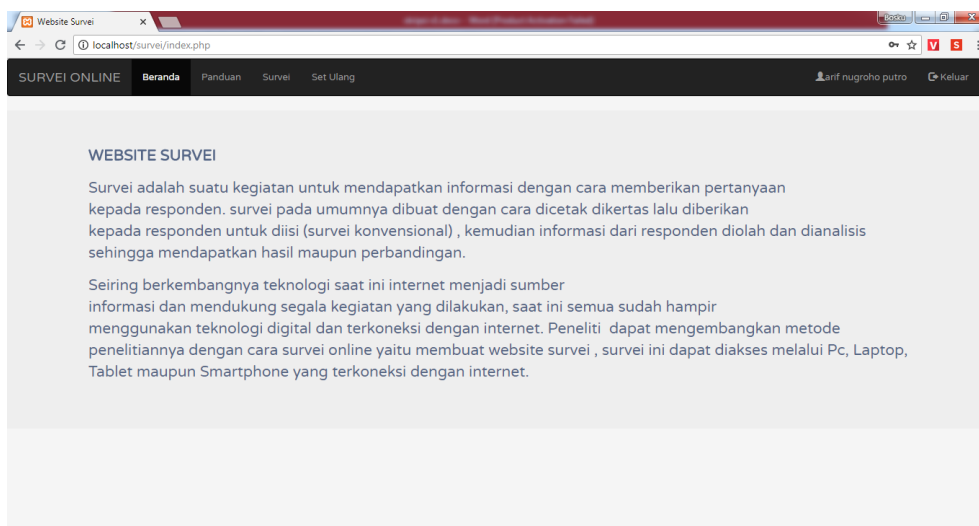
Gambar 1. Halaman Login



Gambar 2. Halaman Pendaftaran

3.1.2 HALAMAN UTAMA (BERANDA)

Gambar 3 merupakan halaman utama pada website ini. halaman ini berisi penjelasan singkat tentang website survei. Pada halaman utama juga terdapat menu beranda, menu panduan, menu survei, menu set ulang, nama responden dan menu keluar. Pada setiap menu yang ada di halaman tersebut memiliki session login, jadi ketika responden ingin mencoba mengakses setiap halaman tersebut tanpa login akan langsung dialihkan ke menu login.

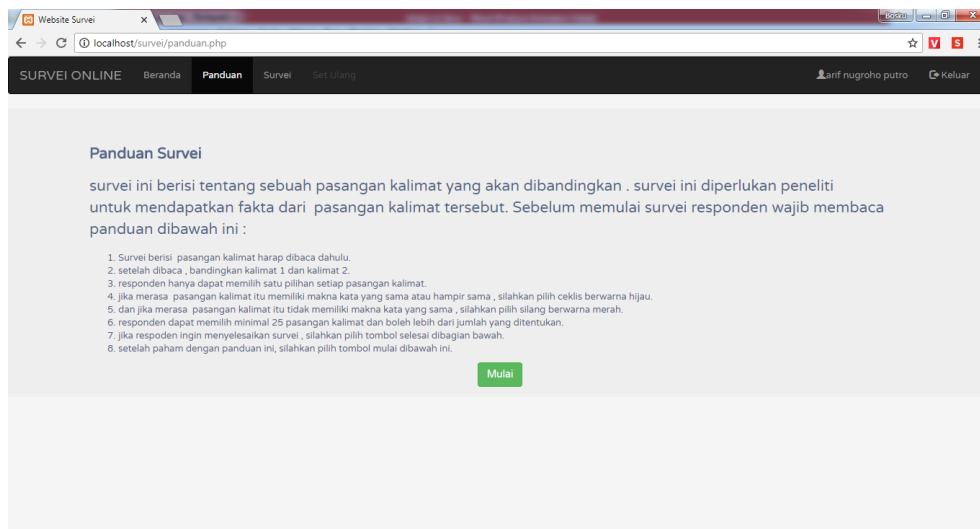


Gambar 3. Halaman Beranda

3.1.3 HALAMAN PANDUAN

Gambar 4 merupakan halaman panduan yang berisi penjelasan dan cara mengisi survei dari website ini. dalam beberapa proses yang dialami peneliti , kebanyakan responden langsung ingin mengisi survei tanpa membaca penjelasan dan cara untuk mengisi survei yang disediakan peneliti, untuk mengatasi hal itu peneliti menyediakan session untuk menu survei didalam menu panduan dengan

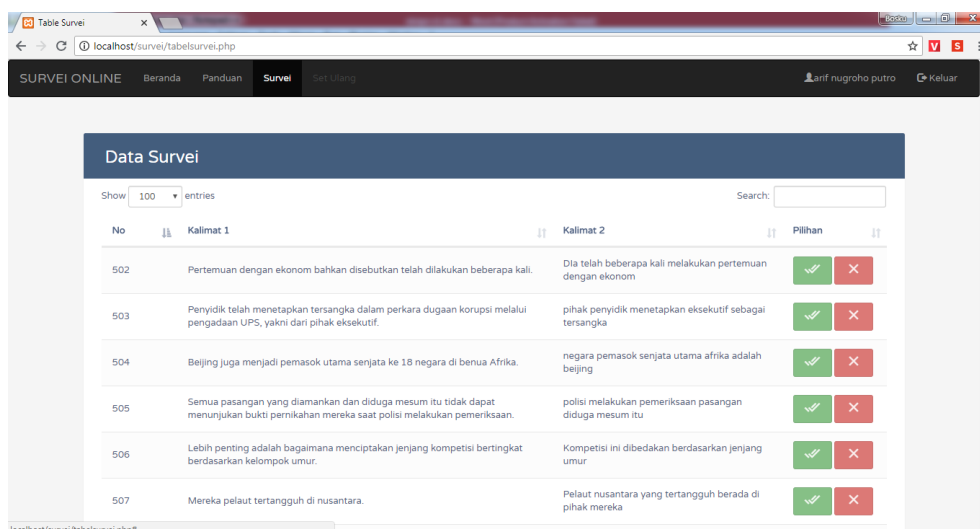
cara menekan tombol mulai, jika responden langsung memilih menu survei, maka responden akan langsung dialihkan ke menu panduan.



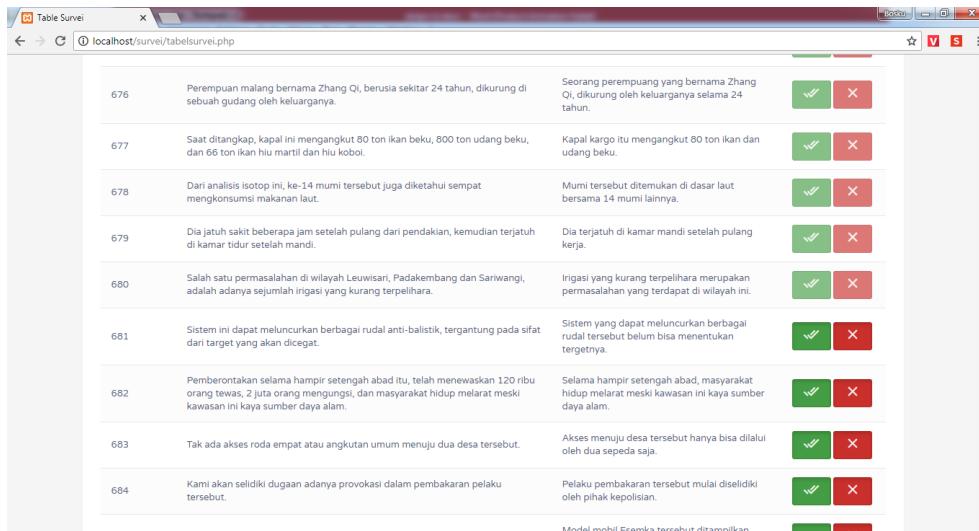
Gambar 4. Halaman Panduan

3.1.4 HALAMAN PENGISIAN SURVEI

Gambar 5 merupakan halaman pengisian survei, pada halaman ini disediakan sebuah tabel yang berisi nomor urut kalimat, kalimat 1 , kalimat 2 , pilihan. responden harus membandingkan pasangan kalimat berdasarkan makna pada penyusun kalimat, jika pasangan kalimat memiliki arti kata yang sama, responden harus memilih tombol centang berwarna hijau dan jika pasangan kalimat memiliki arti kata yang berbeda, responden harus memilih tombol silang berwarna merah. Responden hanya dapat memilih sekali dalam setiap pasangan kalimat, untuk memberikan tanda bahwa responden sudah memilih satu dari pilihan yang disediakan, peneliti menonaktifkan tombol seperti terlihat pada gambar 6 . Halaman ini juga menyediakan tombol untuk mengakhiri survei pada bagian bawah tabel survei ini.



Gambar 5. Halaman Survei

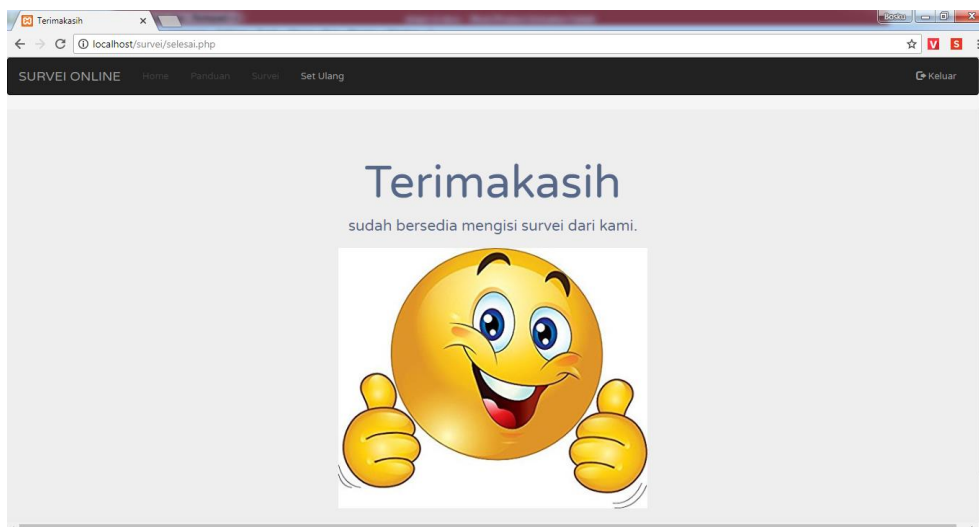


No	Deskripsi	Status
676	Perempuan malang bernama Zhang Qi, berusia sekitar 24 tahun, dikurung di sebuah gudang oleh keluarganya.	Seorang perempuan yang bernama Zhang Qi, dikurung oleh keluarganya selama 24 tahun.
677	Saat ditangkap, kapal ini mengangkut 80 ton ikan beku, 800 ton udang beku, dan 66 ton ikan hiu martil dan hiu koboi.	Kapal kargo itu mengangkut 80 ton ikan dan udang beku.
678	Dari analisis isotop ini, ke-14 mumi tersebut juga diketahui sempat mengonsumsi makanan laut.	Mumi tersebut ditemukan di dasar laut bersama 14 mumi lainnya.
679	Dia jatuh sakit beberapa jam setelah pulang dari pendakian, kemudian terjatuh di kamar tidur setelah mandi.	Dia terjatuh di kamar mandi setelah pulang kerja.
680	Salah satu permasalahan di wilayah Leuwisari, Padakembang dan Sariwangi, adalah adanya sejumlah irigasi yang kurang terpelihara.	Irigasi yang kurang terpelihara merupakan permasalahan yang terdapat di wilayah ini.
681	Sistem ini dapat meluncurkan berbagai rudal anti-balistik, tergantung pada sifat dari target yang akan dicegat.	Sistem yang dapat meluncurkan berbagai rudal tersebut belum bisa menentukan targetnya.
682	Pemberontakan selama hampir setengah abad itu, telah menewaskan 120 ribu orang tewas, 2 juta orang mengungsi, dan masyarakat hidup melarat meski kawasan ini kaya sumber daya alam.	Selama hampir setengah abad, masyarakat hidup melarat meski kawasan ini kaya sumber daya alam.
683	Tak ada akses roda empat atau angkutan umum menuju dua desa tersebut.	Akses menuju desa tersebut hanya bisa dilalui oleh dua sepeda saja.
684	Kami akan selidiki dugaan adanya provokasi dalam pembakaran pelaku tersebut.	Pelaku pembakaran tersebut mulai diselidiki oleh pihak kepolisian.

Gambar 6. Halaman Survei

3.1.5 HALAMAN SELESAI

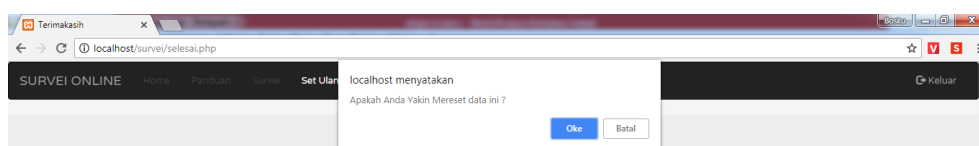
halaman ini merupakan tampilan ketika responden mengakhiri survei pada website ini. seperti pada gambar 7 tampilan ini berupa ucapan terimakasih untuk responden karena sudah bersedia mengisi survei.



Gambar 7. Halaman Selesai

3.1.6 MENU SET ULANG

Gambar 8 adalah menu set ulang yang berfungsi untuk menghilangkan tanda pada tombol yang dinonaktifkan saat di pilih oleh responden. Saat diklik menu set ulang terdapat konfirmasi bahwa akan mereset ulang tombol pilihan survei tersebut.



Gambar 8. Menu Set Ulang

3.2 HASIL

Survei dari pasangan kalimat telah dilakukan sebanyak 1000 pasangan kalimat. Setiap pasangan kalimat dinilai oleh dua responden yang berbeda. Setiap pasangan kalimat juga dilakukan perhitungan kemiripan menggunakan fungsi similaritas.

Dari 1000 pasangan kalimat yang diuji menggunakan website, ada 40 responden yang menilai pasangan kalimat dengan jawaban yang berbeda beda. Kemudian data tersebut diseleksi berdasarkan jawaban sama sehingga mendapatkan data yang berjumlah 660 pasangan kalimat. seperti yang terlihat pada tabel 1. Pada kolom 1 merupakan id pasangan kalimat, kolom 2 dan kolom 3 merupakan pasangan kalimat yang dibandingkan oleh responden dan kolom 4 merupakan skor menurut persepsi responden.

Tabel 1. Tabel skor berdasarkan persepsi seseorang

Id pasangan	Kalimat 1	Kalimat 2	Skor
1	Sehingga semua pelayanan bisa dilakukan di klinik atau rumah sakit langsung tanpa menggunakan surat rujukan.	Semua pelayanan dapat dilakukan di rumah sakit kelas B setelah mendapat rujukan dari dokter keluarga.	Tidak Mirip
2	Ia menjelaskan, apa yang terjadi di Jakarta sebenarnya banyak terjadi di daerah lain.	Kejadian pencurian dengan pemberatan sering terjadi di kampung itu.	Tidak Mirip
3	Acara diisi pula dengan konser amal bertema 'Memperkuat Peran dan Posisi Perempuan untuk Menyambut Hari Perempuan Internasional.	Hari Perempuan Internasional diperingati dengan konser amal.	Mirip
4	Anda hanya akan menggunakan putih telur dalam perawatan ini.	Anda memerlukan putih telur untuk menjalani perawatan ini.	Mirip
5	Mereka yang mencetak gol pada detik-detik terakhir akan merayakannya.	Gol yang dilesakkan pemain bintang itu menyudahi laga final piala dunia.	Mirip

Pada tabel 2. Menampilkan hasil perhitungan dari pasangan kalimat dengan rumus fungsi similaritas. Pada kolom 1 merupakan id pasangan kalimat yang sudah diseleksi, kolom 2 merupakan hasil perhitungan cosinus, kolom 3 hasil perhitungan dice dan kolom 4 hasil perhitungan euclidean. Dari tabel 2. Hasil perhitungan cosinus dan dice hampir sama, sedangkan hasil perhitungan euclidean terlihat lebih besar dari keduanya.

Tabel 2. Tabel perhitungan berdasarkan fungsi similaritas

Id pasangan	Perhitungan cosinus	Perhitungan dice	Perhitungan euclidean
1	0.466667	0.466667	4
2	0.323381	0.363636	4.2426406871193
3	0.520266	0.5	3.7416573867739
4	0.589256	0.588235	2.6457513110646
5	0.201008	0.2	4

Pada tabel 3. Peneliti memberi perbandingan dengan ketentuan pada setiap fungsi similaritas, untuk fungsi cosinus dan fungsi dice memiliki ketentuan yang sama yaitu jika penilaian fungsi cosinus atau dice lebih besar atau sama dengan 0.5 maka pasangan kalimat dikatakan mirip dan jika penilaian fungsi cosinus atau dice kurang dari 0.5 maka pasangan kalimat dikatakan tidak mirip. Sedangkan untuk fungsi euclidean distance, jika penilaian fungsi euclidean kurang dari atau sama dengan 2 maka pasangan kalimat dikatakan mirip dan jika penilaian fungsi euclidean lebih besar dari 2 maka pasangan kalimat dikatakan tidak mirip. Hasil pada setiap fungsi ditampilkan Seperti pada kolom 2 sampai kolom 4.

Tabel 3. Tabel perbandingan menurut ketentuan setiap fungsi similaritas

Id pasangan	Penilaian cosinus	Penilaian dice	Penilaian euclidean
1	Tidak Mirip	Tidak Mirip	Tidak Mirip
2	Tidak Mirip	Tidak Mirip	Tidak Mirip
3	Mirip	Mirip	Tidak Mirip
4	Mirip	Mirip	Tidak Mirip
5	Tidak Mirip	Tidak Mirip	Tidak Mirip

Kemudian hasil persepsi manusia dibandingkan dengan hasil fungsi similaritas Seperti pada Tabel 4, jika hasil persepsi manusia dan hasil similaritas sama akan bernilai 1 dan jika hasil persepsi manusia dan hasil similaritas berbeda akan bernilai 0. Hasilnya pada Tabel 5. kolom 2 sampai kolom 4.

Tabel 4. Tabel Perbandingan Persepsi manusia dengan hasil fungsi similaritas

Id pasangan	Persepsi Seseorang	Penilaian cosinus	Penilaian dice	Penilaian euclidean
1	Tidak Mirip	Tidak Mirip	Tidak Mirip	Tidak Mirip
2	Tidak Mirip	Tidak Mirip	Tidak Mirip	Tidak Mirip
3	Mirip	Mirip	Mirip	Tidak Mirip
4	Mirip	Mirip	Mirip	Tidak Mirip
5	Mirip	Tidak Mirip	Tidak Mirip	Tidak Mirip

Tabel 5. Tabel penilaian persepsi manusia dan fungsi similaritas

Penilaian kesamaan persepsi manusia dan fungsi similaritas			
Id pasangan	Persepsi \equiv cosinus	Persepsi \equiv dice	Persepsi \equiv euclidean
1	1	1	1
2	1	1	1
3	1	1	0
4	1	1	0
5	0	0	0

Dari hasil penilaian kesamaan persepsi manusia dan hasil penilaian fungsi similaritas diatas. Maka dihitung keseluruhan kesamaan setiap fungsi similaritas, hasilnya pada tabel 6 baris 1,2 dan 3, kemudian dibagi keseluruhan data yang diseleksi yaitu 660 pasangan kalimat.

$$C = \frac{\sum A}{\sum B} \quad (4)$$

Keterangan :

C = Skor kesamaan Fungsi Similaritas

$\sum A$ = Jumlah semua data penilaian kesamaan persepsi seseorang \equiv fungsi similaritas

$\sum B$ = Jumlah Semua data persepsi seseorang yang sudah diseleksi

Skor kesamaan *cosine similarity* adalah 0.680303. skor kesamaan *dice similarity* adalah 0.684848. sedangkan skor kesamaan *euclidean distance* adalah 0.348485.

Tabel 6. Hasil perhitungan kesamaan persepsi manusia dan fungsi similaritas

	Persepsi \equiv cosinus	Persepsi \equiv dice	Persepsi \equiv euclidean
Jumlah kesamaan	449	452	230
Hasil skor	0.680303	0.684848	0.348485

4. PENUTUP

Berdasarkan skor persamaan antara hasil seleksi persepsi manusia dan penilaian fungsi similaritas pada pasangan kalimat dapat disimpulkan bahwa dari ketiga fungsi similaritas, fungsi *dice similarity* memiliki skor kesamaan yang terbaik untuk menghitung similaritas kalimat yaitu 0.684848. sedangkan fungsi *euclidean distance* memiliki skor kesamaan yang kurang baik untuk menghitung similaritas kalimat.

DAFTAR PUSTAKA

- Hamzah, A., Soesianto, F., Susanto, A., & Istiyanto, J. E. (2008). Studi Kinerja Fungsi-Fungsi Jarak dan Similaritas dalam Clustering Dokumen Teks Berbahasa Indonesia. *Seminar Nasional Informatika*, 21-30. doi:1979-2328
- Leonardo, B., & Hansun, S. (2017). Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 5(2), 462-471. doi:10.11591/ijeecs.v5.i2.pp462-471
- Putung, K. D., Lumenta, A., & Jacobus, A. (2016). Penerapan Sistem Temu Kembali Informasi pada Kumpulan Dokumen Skripsi. *E-journal Teknik Informatika*, 8(1), 18-23. doi:2301-8364
- Syauqi, A., & Nurwahdah, A. (2015). Sistem Tanya Jawab dengan Web Semantik. *Seminar Nasional Aplikasi Teknologi Informasi*. doi:1907-5022
- Thamrin, H., & Sabardilla, A. (2015). Efektivitas Algoritma Semantik dengan Keterkaitan Kata dalam Mengukur Kemiripan Teks Bahasa Indonesia. *Khazanah Informatika*, 1(1), 7-11. doi:2477-698X
- Thamrin, H., & Wantoro, J. (2014). An Attempt to Create an Automatic Scoring Tool of Short Text Answer in Bahasa Indonesia. *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics*, (pp. 96-98). Yogyakarta.
- Wahyuni, R. T., Prastiyanto, D., & supraptono, E. (2017). Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro*, 9(1), 18-23. doi:1411-0059
- Zamzami, D., Rahutomo, F., & Puspitasari, D. (2016). Aplikasi Wordnet Indonesia Berdasarkan Kamus Thesaurus Bahasa Indonesia Menggunakan Algoritma Rule Based Text Parsing.